HHS Public Access

Author manuscript

Genet Epidemiol. Author manuscript; available in PMC 2018 July 01.

Published in final edited form as:

Genet Epidemiol. 2017 July ; 41(5): 427–436. doi:10.1002/gepi.22046.

Conditional Analysis of Multiple Quantitative Traits Based on Marginal GWAS Summary Statistics

Yangqing Deng[1] and **Wei Pan**[1,2]

[1]Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, USA

Abstract

There has been an increasing interest in joint association testing of multiple traits for possible pleiotropic effects. However, even in the presence of pleiotropy, most of the existing methods cannot distinguish direct and indirect effects of a genetic variant, say SNP, on multiple traits, and a conditional analysis of a trait adjusting for other traits is perhaps the simplest and most common approach to addressing this question. However, without individual-level genotypic and phenotypic data but with only GWAS summary statistics, as typical with most large-scale GWAS consortium studies, we are not aware of any existing method for such a conditional analysis. We propose such a conditional analysis, offering formulas of necessary calculations to fit a joint linear regression model for multiple quantitative traits. Furthermore, our method can also accommodate conditional analysis on multiple SNPs in addition to on multiple quantitative traits, which is expected to be useful for fine mapping. We provide numerical examples based on both simulated and real GWAS data to demonstrate the effectiveness of our proposed approach, and illustrate possible usefulness of conditional analysis by contrasting its result differences from those of standard marginal analyses.

1 Introduction

There is an increasing interest in association analysis of multiple traits with many new tests being recently proposed; see, e.g. He et al. (2013), Jiang et al. (2014), Zhang et al. (2014), Wang et al. (2016), Kim et al. (2016) and Schaid et al. (2016) and references therein. There are two main reasons: one is to increase statistical power and the other is to detect pleotropic effects, which may shed light on underlying biology and for possible repurposing the use of existing drugs. However, as pointed out by Schaid et al. (2016), almost all existing methods test with a null hypothesis of no associated trait, which may be rejected even in the absence of pleiotropy if only one of the multiple traits is indeed associated; they cannot tell whether there is indeed pleiotropy. Accordingly, several methods have appeared to explore which of the multiple traits are indeed associated (Stephens, 2013; Majumdar et al., 2016). In particular, Schaid et al. (2016) proposed a formal testing framework to sequentially test for the number of associated traits. However, even in the presence of multiple associated traits, these approaches cannot distinguish direct and indirect associations. For example, if it is known two traits are associated with one SNP, it is unknown whether one of the traits is a

[2]Corresponding author: weip@biostat.umn.edu.

mediator between the SNP and the other trait, the answer to which would be of interest. For this purpose, a conditional analysis would be useful: we can test for possible association between a trait and an SNP by conditioning on all other traits; if the effect of the trait being tested is indirect and solely through some other trait as a mediator, it will not be significantly associated with the SNP in the conditional analysis, though it may be marginally associated with the SNP (without adjusting for other traits) in a standard GWAS analysis. With the availability of individual-level genotypic and phenotypic data, such a conditional analysis is straightforward. However, due to confidentiality concerns or logistic reasons, often we only have summary statistics available from a GWAS or a meta-analysis of multiple GWAS. Most existing association testing methods for multiple traits are applicable only to individual-level data, though a few exceptions exist for marginal analyses (Zhu et al., 2015; Kim et al., 2015; Cichonska et al., 2016; Kwak, & Pan, 2016, 2017). A major contribution here is to extend a conditional analysis of multiple quantitative traits to GWAS summary statistics without individual-level data. Our idea is similar to conditional analysis of a single quantitative trait on multiple SNPs with only GWAS summary statistics (Yang et al., 2012). The difference is that, in our approach we use the GWAS summary statistics to estimate the correlations among the multiple quantitative traits, while in the latter a reference panel of individual-level genotypic data is used to estimate the correlations (LD) among the SNPs. We will also combine the two approaches for a joint conditional analysis of multiple traits and multiple SNPs. The proposed approach will be useful to sort out specific effects of the SNPs in a locus associated with one or more traits, e.g. in fine mapping. It can be also used to infer the causal relationships among multiple quantitative traits as in structural equation modeling (Li et al., 2006).

A caveat of our approach is that we implicitly assume that a set of multiple traits is collected on each subject, as in usual multivariate analyses. However, for GWAS summary statistics, this assumption may not exactly hold due to missing values of one or more traits on some subjects. This can be easily seen from unequal numbers of subjects across the multiple traits for the same SNP in a typical dataset of GWAS summary statistics. The consequence is biased parameter estimation with the bias increasing with the proportion of non-overlapping subjects across the traits. In the extreme case that no more than one trait is collected from any subject, then the multiple traits will be estimated as uncorrelated, implying that a conditional analysis of multiple traits will be equal to marginal analyses on each trait separately. We will use simulations to demonstrate this phenomenon. Finally, we will apply our proposed method to two large GWAS datasets, confirming differences between usual marginal analyses and a conditional analysis.

2 Methods

2.1 Adjusting for One Phenotype

We first consider the simplest case with two traits and a single SNP. Denote the two traits as $\mathbf{Y}_1 = (y_{i1})_{n\times 1}$ and $\mathbf{Y}_2 = (y_{i2})_{n\times 1}$, and the additive coding of the SNP as $\mathbf{X}_1 = (x_{i1})_{n\times 1}$, where n is the number of subjects. We assume that both the two traits and the SNP have been centered so that an intercept is not needed in a regression analysis. Given a dataset of GWAS

summary statistics, we have the estimates $\widehat{\beta_1}$, $\widehat{\beta_2}$ and $\widehat{\mathrm{var}}(\widehat{\beta_1})$, $\widehat{\mathrm{var}}(\widehat{\beta_2})$ from a marginal analysis with marginal linear models

$$\boldsymbol{Y}_1=\mathbf{X}_1\beta_1+\mathbf{e}_1$$

$$\boldsymbol{Y}_2=\mathbf{X}_1\beta_2+\mathbf{e}_2$$

$$\mathbf{e}_i\sim\mathrm{N}(\mathbf{0},{\sigma_i}^2\mathbf{I}_n)\quad i=1,2$$

Now our goal is to obtain parameter estimates from a conditional analysis with a joint linear regression model

$$\boldsymbol{Y}_1=(\mathbf{X}_1\quad\boldsymbol{Y}_2)\begin{pmatrix}b_1\\b_2\end{pmatrix}+\mathbf{e}$$

where b_1, b_2 are the joint effects, and $\mathbf{e}\sim\mathrm{N}(\mathbf{0},\sigma^2\mathbf{I}_n)$. In particular, b_1 is the parameter of interest, giving conditional association of the SNP with trait 1 after adjusting for trait 2. Based on the ordinary least squares estimator, we have

$$\begin{pmatrix}\widehat{b_1}\\\widehat{b_2}\end{pmatrix}=\begin{pmatrix}\mathbf{X}_1'\mathbf{X}_1 & \boldsymbol{Y}_2'\mathbf{X}_1\\\mathbf{X}_1'\boldsymbol{Y}_2 & \boldsymbol{Y}_2'\boldsymbol{Y}_2\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{X}_1'\boldsymbol{Y}_1\\\boldsymbol{Y}_2'\boldsymbol{Y}_1\end{pmatrix}$$

$$\widehat{\mathrm{var}}\begin{pmatrix}\widehat{b_1}\\\widehat{b_2}\end{pmatrix}=\begin{pmatrix}\mathbf{X}_1'\mathbf{X}_1 & \boldsymbol{Y}_2'\mathbf{X}_1\\\mathbf{X}_1'\boldsymbol{Y}_2 & \boldsymbol{Y}_2'\boldsymbol{Y}_2\end{pmatrix}^{-1}{\hat{\sigma}_J}^2$$

where $\hat{\sigma}_J^2$ is the estimated residual variance for the joint model.

$\hat{\sigma}_J^2$ can be calculated as

$${\hat{\sigma}_J}^2=\frac{\boldsymbol{Y}_1'\boldsymbol{Y}_1-\begin{pmatrix}\widehat{b_1}\\\widehat{b_2}\end{pmatrix}'\begin{pmatrix}\mathbf{X}_1'\boldsymbol{Y}_1\\\boldsymbol{Y}_2'\boldsymbol{Y}_1\end{pmatrix}}{n-2}$$

Hence, we only need $\mathbf{X}_1'\mathbf{X}_1$, $\mathbf{X}_1'\mathbf{Y}_1$, $\mathbf{X}_1'\mathbf{Y}_2$, $\mathbf{Y}_1'\mathbf{Y}_1$, $\mathbf{Y}_2'\mathbf{Y}_2$ and $\mathbf{Y}_1'\mathbf{Y}_2$ to estimate the coefficients and their covariance matrix.

We can obtain $\mathbf{X}_1'\mathbf{Y}_1$, $\mathbf{X}_1'\mathbf{Y}_2$, $\mathbf{Y}_1'\mathbf{Y}_1$ and $\mathbf{Y}_2'\mathbf{Y}_2$ using the following equations

$$\widehat{\beta_1}=\frac{\mathbf{X}_1'\boldsymbol{Y}_1}{\mathbf{X}_1{}'\mathbf{X}_1}$$

$$\widehat{\beta_2}=\frac{\mathbf{X}_1'\boldsymbol{Y}_2}{\mathbf{X}_1{}'\mathbf{X}_1}$$

$$\widehat{\sigma_1}^2=\frac{\boldsymbol{Y}_1'\boldsymbol{Y}_1-\mathbf{X}_1'\mathbf{X}_1\widehat{\beta_1}^2}{n-1}=\mathbf{X}_1'\mathbf{X}_1\widehat{\mathrm{var}}(\widehat{\beta_1})$$

$$\widehat{\sigma_2}^2=\frac{\boldsymbol{Y}_2'\boldsymbol{Y}_2-\mathbf{X}_1'\mathbf{X}_1\widehat{\beta_2}^2}{n-1}=\mathbf{X}_1'\mathbf{X}_1\widehat{\mathrm{var}}(\widehat{\beta_2})$$

where $\widehat{\sigma_1}$ and $\widehat{\sigma_2}$ are the estimated residual variances for the two marginal models.

Suppose $\mathbf{X}_2$, $\mathbf{X}_3$, … $\mathbf{X}_{m+1}$ are some other SNPs that do not have marginal associations with both $\mathbf{Y}_1$ and $\mathbf{Y}_2$. Following Kim et al. (2015), if we have the Z-scores $\mathbf{Z}_1 = (Z_{i1})_{m\times 1}$ for β_1 in the marginal models $\mathbf{Y}_1 = \mathbf{X}_{i+1}\beta_1 + \mathbf{e_1}$, and $\mathbf{Z}_2 = (Z_{i2})_{m\times 1}$ for $\mathbf{Y}_2 = \mathbf{X}_{i+1}\beta_2 + \mathbf{e_2}$, and if the sample size is large (as usual in GWAS), we can estimate $\mathbf{Y}_1{}'\mathbf{Y}_2$ as $\sqrt{\boldsymbol{Y}_1'\boldsymbol{Y}_1\boldsymbol{Y}_2'\boldsymbol{Y}_2}\mathrm{cor}(\boldsymbol{Z}_1,\boldsymbol{Z}_2)$, since we have

$$\frac{\boldsymbol{Y}_1'\boldsymbol{Y}_2}{\sqrt{\boldsymbol{Y}_1'\boldsymbol{Y}_1\boldsymbol{Y}_2'\boldsymbol{Y}_2}}\approx\mathrm{cor}(\boldsymbol{Y}_1,\boldsymbol{Y}_2)\approx\mathrm{cor}(\boldsymbol{Z}_1,\boldsymbol{Z}_2)$$

where cor($\mathbf{Z}_1$, $\mathbf{Z}_2$) is the sample Pearson correlation between the two sets of the Z-statistics $\mathbf{Z}_1$ and $\mathbf{Z}_2$.

When the summary statistics are fixed, the calculated $\mathbf{X}_1{}'\mathbf{Y}_1$, $\mathbf{X}_1{}'\mathbf{Y}_2$, $\mathbf{Y}_1{}'\mathbf{Y}_1$ and $\mathbf{Y}_2{}'\mathbf{Y}_2$ are proportional to $\mathbf{X}_1{}'\mathbf{X}_1$. As a result, $\mathbf{Y}_1{}'\mathbf{Y}_2$ is also proportional to $\mathbf{X}_1{}'\mathbf{X}_1$, and the calculated coefficients and covariance do not depend on the value of $\mathbf{X}_1{}'\mathbf{X}_1$ we specify. Hence, we can simply set $\mathbf{X}_1{}'\mathbf{X}_1$ to 1.

As a result, we have

$$\begin{pmatrix}\widehat{b_1}\\ \widehat{b_2}\end{pmatrix}=\begin{pmatrix}1 & \widehat{\beta_2}\\ \widehat{\beta_2} & S_2\end{pmatrix}^{-1}\begin{pmatrix}\widehat{\beta_1}\\ \mathrm{cor}(\boldsymbol{Z}_1,\boldsymbol{Z}_2)\sqrt{S_1S_2}\end{pmatrix}$$

$$\hat{\sigma}_J^2 = \frac{S_1 - \begin{pmatrix} \widehat{b_1} \\ \widehat{b_2} \end{pmatrix}' \begin{pmatrix} S_1 \\ \mathrm{cor}(\boldsymbol{Z}_1, \boldsymbol{Z}_2)\sqrt{S_1 S_2} \end{pmatrix}}{n-2}$$

$$\widehat{\mathrm{var}} \begin{pmatrix} \widehat{b_1} \\ \widehat{b_2} \end{pmatrix} = \begin{pmatrix} 1 & \widehat{\beta_2} \\ \widehat{\beta_2} & S_2 \end{pmatrix}^{-1} \hat{\sigma}_J^2$$

where $S_1 = (n-1)\widehat{\mathrm{var}}(\widehat{\beta_1}) + \widehat{\beta_1}^2$, $S_2 = (n-1)\widehat{\mathrm{var}}(\widehat{\beta_2}) + \widehat{\beta_2}^2$. Again $\mathbf{Z}_1$, $\mathbf{Z}_2$ are the Z-scores for a large number, say a few hundreds of thousands, of some null SNPs' for $\mathbf{Y}_1$, $\mathbf{Y}_2$ respectively, drawn from the given GWAS summary statistics.

We note that in practice, the subjects used to generate summary statistics for $\mathbf{Y}_1$ and $\mathbf{Y}_2$ may not be exactly the same, which may lead to biased estimates; the degree of the bias increases with the proportion of the subjects who did not have all the measurements for all the traits. For operational purpose, we can set n as the number of all subjects and carry on the above process.

2.2 Adjusting for Multiple Phenotypes and Multiple SNPs

We extend our method to the general situation where $\mathbf{Y}_2$ includes more than one phenotype, and $\mathbf{X}_1$ includes more than one SNP. Suppose $\mathbf{Y}_2 = (y_{ik})_{n\times(K-1)}$, $k = 2, \ldots K$, where $(K-1)$ is the number of phenotypes we want to adjust for in the joint model; $\mathbf{X}_1 = (x_{il})_{n\times L}$, where L is the number of SNPs. Based on the given GWAS data, we have the summary statistics $\widehat{\beta_{\mathrm{kl}}}$ and $\widehat{\mathrm{var}}(\widehat{\beta_{\mathrm{kl}}})$ in marginal analyses with marginal linear models

$$\boldsymbol{Y}_{(k)} = \mathbf{X}_{1l}\beta_{\mathrm{kl}} + \mathbf{e}_{\mathrm{kl}},$$

where $\mathbf{Y}_{(k)}$ is the kth column of $(\mathbf{Y}_1\ \mathbf{Y}_2)$, $k = 1, \ldots K$, $\mathbf{X}_{1l} = (x_{il})_{n\times 1}$, and $\mathbf{e}_{kl} \sim N(\mathbf{0}, \sigma_{kl}^2 \mathbf{I}_n)$. Now we are interested in inference for conditional analysis with a joint (full) linear regression model

$$\boldsymbol{Y}_1 = (\mathbf{X}_1\ \ \boldsymbol{Y}_2) \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} + \mathbf{e}$$

where $\mathbf{b_1}$, $\mathbf{b_2}$ are the joint effects, and $\mathbf{e} \sim N(\mathbf{0}, \sigma_f^2 \mathbf{I}_n)$.

In practice, due to missing values, different summary statistics often came from different numbers of individuals. Suppose n_{kl} is the sample size used to calculate $\widehat{\beta_{\mathrm{kl}}}$ and $\widehat{\mathrm{var}}(\widehat{\beta_{\mathrm{kl}}})$, and n is the number of all subjects; we assume $n = \max_{kl} n_{\mathrm{kl}}$. Each sample of $n_{kl} < n$ individuals is regarded as a random sample from the n subjects.

Similar to what we did before, we have

$$\begin{pmatrix} \widehat{\mathbf{b_1}} \\ \widehat{\mathbf{b_2}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\boldsymbol{Y}_2 \\ \boldsymbol{Y}_2'\mathbf{X}_1 & \boldsymbol{Y}_2'\boldsymbol{Y}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_1'\boldsymbol{Y}_1 \\ \boldsymbol{Y}_2'\boldsymbol{Y}_1 \end{pmatrix}$$

$$\widehat{\mathrm{var}} \begin{pmatrix} \widehat{\mathbf{b_1}} \\ \widehat{\mathbf{b_2}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\boldsymbol{Y}_2 \\ \boldsymbol{Y}_2'\mathbf{X}_1 & \boldsymbol{Y}_2'\boldsymbol{Y}_2 \end{pmatrix}^{-1} \hat{\sigma}_J{}^2$$

The residual variance $\hat{\sigma}_J^2$ can be estimated as

$$\hat{\sigma}_J{}^2 = \frac{\boldsymbol{Y}_1'\boldsymbol{Y}_1 - \begin{pmatrix} \widehat{\mathbf{b_1}} \\ \widehat{\mathbf{b_2}} \end{pmatrix}' \begin{pmatrix} \mathbf{X}_1'\boldsymbol{Y}_1 \\ \boldsymbol{Y}_2'\boldsymbol{Y}_1 \end{pmatrix}}{n - (K + L - 1)}$$

Denote the $(k-1)$th column of $\mathbf{Y}_2$ by $\mathbf{Y}_{(k)}$ and $\mathbf{Y}_1$ by $\mathbf{Y}_{(1)}$. We have

$$\boldsymbol{Y}_2 = (\boldsymbol{Y}_{(2)} \quad \ldots \quad \boldsymbol{Y}_{(K)})$$

$$\boldsymbol{Y}_2'\boldsymbol{Y}_2 = (\boldsymbol{Y}_{(k)}'\boldsymbol{Y}_{(q)})_{k=2,\ldots,K;q=2,\ldots,K}$$

$$\mathbf{X}_1'\mathbf{X}_1 = (\mathbf{X}_{1i}'\mathbf{X}_{1j})_{L\times L}$$

$$\mathbf{X}_1'\boldsymbol{Y}_2 = (\mathbf{X}_{1i}'\boldsymbol{Y}_{(k)})_{i=1,\ldots,L;k=2,\ldots,K}$$

$$\boldsymbol{Y}_1'\boldsymbol{Y}_2 = (\boldsymbol{Y}_1'\boldsymbol{Y}_{(k)})_{k=2,\ldots,K}$$

We can obtain $\mathbf{X}_{1l}{}'\mathbf{Y}_{(k)}$, $\mathbf{Y}_{(k)}{}'\mathbf{Y}_{(k)}$ using

$$\widehat{\beta_{\mathrm{kl}}}' = \frac{\mathbf{X}_{1l}'\boldsymbol{Y}_{(k)}}{\mathbf{X}_{1l}'\mathbf{X}_{1l}}$$

$$\frac{\boldsymbol{Y}_{(k)}'\boldsymbol{Y}_{(k)} - \mathbf{X}_{11}'\mathbf{X}_{11}\widehat{\beta_{k1}}'^2}{n-1} = \mathbf{X}_{11}'\mathbf{X}_{11}\widehat{\mathrm{var}}(\widehat{\beta_{k1}}')$$

Note that $\widehat{\beta_{\mathrm{kl}}}'$ is the estimated coefficient using all n subjects. It is different from the $\widehat{\beta_{\mathrm{kl}}}$ we have, which only used n_{kl} subjects. We can replace the unknown $\widehat{\beta_{\mathrm{kl}}}'$ by $\widehat{\beta_{\mathrm{kl}}}$, assuming the sample size is large enough (which is the case in GWAS). As for $\widehat{\mathrm{var}}(\widehat{\beta_{k1}}')$, we should replace it by $n_{k1}\widehat{\mathrm{var}}(\widehat{\beta_{k1}})/n$ because the sample size does influence the variance of the estimator. Then we have

$$\widehat{\beta_{\mathrm{kl}}}=\frac{\mathbf{X}_{1l}'\boldsymbol{Y}_{(k)}}{\mathbf{X}_{1l}'\mathbf{X}_{1l}}$$

$$\frac{\boldsymbol{Y}_{(k)}'\boldsymbol{Y}_{(k)}-\mathbf{X}_{11}'\mathbf{X}_{11}\widehat{\beta_{k1}}^2}{n-1}=\frac{n_{k1}}{n}\mathbf{X}_{11}'\mathbf{X}_{11}\widehat{\mathrm{var}}(\widehat{\beta_{k1}})$$

When n is large, we can also use

$$\boldsymbol{Y}_{(k)}'\boldsymbol{Y}_{(k)}-\mathbf{X}_{11}'\mathbf{X}_{11}\widehat{\beta_{k1}}^2\approx n_{k1}\mathbf{X}_{11}'\mathbf{X}_{11}\widehat{\mathrm{var}}(\widehat{\beta_{k1}})$$

As before, given the GWAS data of summary statistics, we have Z-statistics $\mathbf{Z}_k$ for $\mathbf{Y}_{(k)}$ versus SNPs that are not significant for all phenotypes, which are used to estimate $\mathbf{Y}_{(k)}{}'\mathbf{Y}_{(q)}$ as $\sqrt{\boldsymbol{Y}_{(k)}'\boldsymbol{Y}_{(k)}\boldsymbol{Y}_{(q)}'\boldsymbol{Y}_{(q)}}\mathrm{cor}(\boldsymbol{Z}_k,\boldsymbol{Z}_q)$ when $k\neq q$.

If we can get individual level data $\tilde{\mathbf{X}}_1$ from some reference panel for the L SNPs of interest, we can estimate $\mathbf{X}_1{}'\mathbf{X}_1$ as Σ, where Σ is the sample variance-covariance matrix obtained from $\tilde{\mathbf{X}}_1$.

To summarize, if we have $\widehat{\beta_{\mathrm{kl}}}$ and $\widehat{\mathrm{var}}(\widehat{\beta_{\mathrm{kl}}})$ for $\mathbf{Y}_{(k)}\sim\mathbf{X}_{1l}$ ($k=1\ldots K$, $l=1\ldots L$) and n_{kl}, the sample size to get the summary statistics, as well as Σ, an estimate of the variance-covariance matrix of the SNPs, and $\mathbf{Z}_k$, Z-scores for $\mathbf{Y}_{(k)}$ versus null SNPs, we can use the following procedure to get the coefficients and the variance estimates:

$$\mathbf{X}_1'\mathbf{X}_1=\sum=(\mathrm{s_{ij}})_{L\times L}$$

$$\mathbf{X}_1'\boldsymbol{Y}_1=\begin{pmatrix}s_{11}&0&\cdots&0\\0&s_{22}&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&s_{\mathrm{LL}}\end{pmatrix}(\widehat{\beta_{11}}\widehat{\beta_{12}}\ldots\widehat{\beta_{1L}})'$$

$$\mathbf{X}_1'\boldsymbol{Y}_2=\begin{pmatrix} s_{11} & 0 & \cdots & 0 \\ 0 & s_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_{\mathrm{LL}} \end{pmatrix}\begin{pmatrix} \widehat{\beta_{21}} & \widehat{\beta_{31}} & \cdots & \widehat{\beta_{K1}} \\ \widehat{\beta_{22}} & \widehat{\beta_{32}} & \cdots & \widehat{\beta_{K2}} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\beta_{2L}} & \widehat{\beta_{3L}} & \cdots & \widehat{\beta_{\mathrm{KL}}} \end{pmatrix}$$

$$\boldsymbol{Y}'_{(k)}\boldsymbol{Y}_{(k)}=n_{k1}s_{11}\widehat{\mathrm{var}}(\widehat{\beta_{k1}})+s_{11}\widehat{\beta_{k1}}^2$$

$$\boldsymbol{Y}'_{(k)}\boldsymbol{Y}_{(q)}=\sqrt{\boldsymbol{Y}'_{(k)}\boldsymbol{Y}_{(k)}\boldsymbol{Y}'_{(q)}\boldsymbol{Y}_{(q)}}\mathrm{cor}(\boldsymbol{Z}_k,\boldsymbol{Z}_q), k\neq q$$

$$\boldsymbol{Y}_1'\boldsymbol{Y}_1=\boldsymbol{Y}'_{(1)}\boldsymbol{Y}_{(1)}$$

$$\boldsymbol{Y}_2'\boldsymbol{Y}_2=(\boldsymbol{Y}'_{(i)}\boldsymbol{Y}_{(j)})_{i=2\ldots K;j=2\ldots K}$$

$$\boldsymbol{Y}_1'\boldsymbol{Y}_2==(\boldsymbol{Y}'_{(1)}\boldsymbol{Y}_{(j)})_{j=2\ldots K}$$

$$\begin{pmatrix} \widehat{\mathbf{b_1}} \\ \widehat{\mathbf{b_2}} \end{pmatrix}=\begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\boldsymbol{Y}_2 \\ \boldsymbol{Y}_2'\mathbf{X}_1 & \boldsymbol{Y}_2'\boldsymbol{Y}_2 \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{X}_1'\boldsymbol{Y}_1 \\ \boldsymbol{Y}_2'\boldsymbol{Y}_1 \end{pmatrix}$$

$$\hat{\sigma}_J{}^2=\frac{\boldsymbol{Y}_1'\boldsymbol{Y}_1-\begin{pmatrix} \widehat{\mathbf{b_1}} \\ \widehat{\mathbf{b_2}} \end{pmatrix}'\begin{pmatrix} \mathbf{X}_1'\boldsymbol{Y}_1 \\ \boldsymbol{Y}_2'\boldsymbol{Y}_1 \end{pmatrix}}{n-(K+L-1)} \quad (n=\max(n_{\mathrm{kl}}))$$

$$\widehat{\mathrm{var}}\begin{pmatrix} \widehat{\mathbf{b_1}} \\ \widehat{\mathbf{b_2}} \end{pmatrix}=\begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\boldsymbol{Y}_2 \\ \boldsymbol{Y}_2'\mathbf{X}_1 & \boldsymbol{Y}_2'\boldsymbol{Y}_2 \end{pmatrix}^{-1}\hat{\sigma}_J{}^2$$

Note that the calculated $(\widehat{\mathbf{b_1}},\widehat{\mathbf{b_2}})'$ does not depend on the choice of n, but $\hat{\sigma}_J^2$ does. When there are many non-overlapping subjects, e.g. if $\min_{k,l}(n_{\mathrm{kl}})/\max_{k,l}(n_{\mathrm{kl}})$ is small, we shall obtain biased estimates of $\hat{\sigma}_J^2$ and other parameters.

From our experience, the estimate of $\mathbf{X}_1{}'\mathbf{X}_1$ tends to affect the result. In some cases, it may even make some diagonal elements of $\widehat{\mathrm{var}}\begin{pmatrix}\widehat{\mathbf{b_1}}\\ \widehat{\mathbf{b_2}}\end{pmatrix}$ negative. To fix this problem, we may use a modified version of $\mathbf{A}=\begin{pmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\boldsymbol{Y}_2\\ \boldsymbol{Y}_2'\mathbf{X}_1 & \boldsymbol{Y}_2'\boldsymbol{Y}_2\end{pmatrix}$ when calculating $\begin{pmatrix}\widehat{\mathbf{b_1}}\\ \widehat{\mathbf{b_2}}\end{pmatrix}$ and $\widehat{\mathrm{var}}\begin{pmatrix}\widehat{\mathbf{b_1}}\\ \widehat{\mathbf{b_2}}\end{pmatrix}$. Inspired by Cichonska et al. (2016), we can use $(\mathbf{A}+\lambda\mathbf{I})/(\lambda+1)$, where $\mathbf{I}$ is an identity matrix and $\lambda > 0$ is a small positive number to be manually selected. In the data analysis part, we still used the default $\lambda = 0$ unless specified otherwise.

2.3 Hypothesis Testing

Denote $(\mathbf{b_1}, \mathbf{b_2})'$ and $(\widehat{\mathbf{b_1}}, \widehat{\mathbf{b_2}})'$ by $\mathbf{b}$ and $\hat{\mathbf{b}}$. Since we are able to obtain $\hat{\mathbf{b}}$ and $\widehat{\mathrm{var}}(\hat{\mathbf{b}})$, we can test any null hypothesis in the form of $\mathbf{Cb} = \mathbf{l}$ by the Wald test, where $\mathbf{C}$ is a rank Q matrix with $(K + L)$ columns and Q rows, $\mathbf{l}$ is a Q dimensional vector. If Q is greater than 1, the test is testing multiple hypothesis, otherwise a single hypothesis. The test statistic is

$$T=\left[\mathbf{C}\begin{pmatrix}\widehat{\mathbf{b_1}}\\ \widehat{\mathbf{b_2}}\end{pmatrix}-\mathbf{l}\right]'\left[\mathbf{C}\widehat{\mathrm{var}}\begin{pmatrix}\widehat{\mathbf{b_1}}\\ \widehat{\mathbf{b_2}}\end{pmatrix}\mathbf{C}'\right]^{-1}\left[\mathbf{C}\begin{pmatrix}\widehat{\mathbf{b_1}}\\ \widehat{\mathbf{b_2}}\end{pmatrix}-\mathbf{l}\right]$$

Under the null hypothesis, T approximately follows a chi-squared distribution with Q degrees of freedom.

3 Results

3.1 Simulations

To evaluate the effectiveness of our proposed approach, we first compared the performance of using summary statistics only with that of using individual-level data in a simulation study with one SNP and two traits. We generated the genotypes of a single SNP from a binomial distribution with frequency f_0. Then we generated $\mathbf{Y}_2$ using the model $\mathbf{Y}_2 = \mathbf{X}_1\beta_2 + \mathbf{e}_1$, and $\mathbf{Y}_1$ using the model $\boldsymbol{Y}_1=(\mathbf{X}_1 \quad \boldsymbol{Y}_2)\begin{pmatrix}b_1\\ b_2\end{pmatrix}+\mathbf{e}_2$, where $\mathbf{e}_i$ follows a standard normal distribution. To obtain the Z-scores, we generated 500 independent null SNPS with varying allele frequencies $f\sim$U [0.08,0.2] and then estimated their marginal effect sizes and variances. The sample size was set to n. We compared our approach using only summary statistics with the golden standard approach of using individual-level data.

Next, we did simulations to assess the influence of non-overlapping subjects. After generating the data for n subjects, we only used the first n_1 subjects to obtain summary statistics for $\mathbf{Y}_1$, and the last n_2 subjects for $\mathbf{Y}_2$. When building the model based on individual data, we only used the overlapping subjects. We obtained the means and the standard deviations using 100 iterations. We also compared the results with marginal estimates.

In addition, we looked at the performance of the new method for adjusting for multiple traits and multiple SNPs. A new simulation was conducted based on a dataset from the Genetic

Investigation of Anthropometric Traits (GIANT) consortium (Randall et al., 2013). The dataset contains summary statistics for the marginal effects of each single SNPs on each of the anthropometric traits like body mass index (BMI), hip circumference (HIP), waist circumference (WC) and weight, etc., stratified by sex.

For this simulation, we only simulated data for males. We chose 8 SNPs that are marginally significant for the four traits (p-value < 5e-8) but not highly correlated. We generated the genotypes from a multivariate binomial distribution based on the 8 SNPs' minor allele frequencies and their correlations estimated from the reference panel. In this step, we multiplied the correlation between each pair of SNPs by 0.8 to avoid numerical problems. We also standardized the genotypes. Then we simulated HIP and WC using the estimated models for HIP~SNP_1 + SNP_2 + SNP_3 + SNP_4, WC~SNP_5 + SNP_6 + SNP_7 + SNP_8. Next, we generated BMI using the estimated model for BMI~HIP + WC + SNP_1 + SNP_2 + ⋯ + SNP_8. The sample size was set to 2000. After that, we conducted the marginal analysis. Then we did the conditional analysis using the obtained summary statistics.

The results shown in Table 4 indicate that the regression coefficient estimates from the conditional model tended to be more accurate than those from the marginal models, and that they had smaller standard errors, suggesting possible power gains from conditional analyses.

3.2 BCX Data

The Blood Cell Consortium (BCX) data (Chami et al., 2016) contains summary statistics for single SNP effects on red blood cell traits. We applied the method to traits red blood cell count (RBC) and hematocrit (HCT) for European subjects, denoted by $\mathbf{Y}_1$ and $\mathbf{Y}_2$. We examined each SNP that has both significant marginal effects on BCX and HCT. The number of such SNPs is 25. We used the Z-scores for those 206751 SNPs in the same chromosome that were not marginally associated with either trait.

Y_2 in every case turned out to be highly significant, while only 15 SNPs (e.g. rs218237, rs172629) remained significant (p-value < 5e-8) after adjusting for Y_2 in conditional analyses. This might suggest that the other SNPs (e.g. rs11611647, rs837763) influence RBC through HCT. Table 4 shows some examples.

3.3 GIANT Data

We applied the methods to the GIANT data. Before conducting the analysis, for each sex, we estimated the correlations among the traits using 2723514 SNPs that were not marginally significant for BMI, HIP, WC and weight (i.e. p-value > 5e-8) for men and 2723477 SNPs for women. The number of subjects used to obtain the summary statistics varies a lot, depending on SNPs, traits and genders, as shown in Table 5.

First, we looked at the data for males. We considered those SNPs that were marginally significant for BMI, HIP, WC and weight (p-value < 5e-8). Only 44 SNPs satisfied this condition, and they were all mapped to gene FTO on chromosome 16. Since these SNPs were highly correlated, we chose 8 of them so that the correlation matrix estimated from the reference panel was not nearly singular. The panel we used was the 381 European subjects from the 1000 Genomes Project data (The 1000 Genomes Project Consortium, 2015). If the

absolute value of the correlation between two SNPs was greater than 0.98, we excluded the second one. We list the effective sample sizes in Table 6. Note that, due to the inclusion of some family-based studies in the meta-analysis, the effective sample sizes may not be integers.

We conducted a conditional analysis with a joint (full) model for BMI vs. the other traits plus the 8 SNPs. We also considered models adjusting for traits only or for SNPs only. The four types of models can be viewed as BMI~SNP_i (marginal), BMI~SNP_i + HIP + WC + Weight (adjusted for only the traits), BMI~$SNP_1 + SNP_2 + \cdots + SNP_8$ (adjusted for only the SNPs) and BMI~HIP + WC + Weight + $SNP_1 + SNP_2 + \cdots + SNP_8$ (adjusted for both the traits and SNPs).

As shown in Figure 1, after adjusting for both SNPs and traits, three marginally significant SNPs became insignificant, while one became more significant (rs8057044). If we only adjusted for traits, all 8 SNPs became insignificant. The estimated effects were all very small. If we only adjusted for SNPs, a few SNPs retained much larger effect sizes. The effect size of rs8057044 was much larger than those of the other SNPs, yielding p-value 0. Note that SNP rs11075987 was insignificant in the full model, but highly significant in the partial model only adjusting for SNPs, which demonstrates the effects of conditional analysis.

Next, we built the joint model with the same SNPs for females. The results are shown in Figure 2. After adjusting for SNPs and traits, 2 SNPs became insignificant. The most significant SNP in our analysis for males, rs8057044, was much less significant for females: the estimated effect sizes were 0.239 and 0.095, with the standard errors 0.013 and 0.011, respectively. In contrast, in the marginal models, the corresponding effect sizes were 0.072 and 0.058 respectively. The joint analysis seems able to better capture the sex difference for this SNP.

Furthermore, we considered both marginally significant and insignificant SNPs mapped to gene *FTO*. We looked at the 8 significant SNPs from above plus 20 SNPs that are not marginally significant for BMI in the male data. Again we selected 20 SNPs so that none of the pairwise correlation was greater than 0.9, while intentionally reserving rs8057044. Then we built the joint models for men and women respectively.

As shown in Figure 3, some SNPs (e.g. rs9922370, rs12447427, rs12596862) that were not marginally significant for association with BMI became highly significant in the joint model for men. However, their estimated effect sizes were not as large in the joint model for women. Note that the reference panel we used was not stratified by sex.

We also only looked at the 17 marginally insignificant SNPs among the selected SNPs. As shown in Figure 4, none of them became significant in the joint model, suggesting that their significant results obtained earlier were due to their correlations with other marginally significant SNPs.

4 Discussion

We have presented the conditional analysis adjusting for one or more traits and possibly for one or more SNPs using only GWAS summary statistics. Our simulation study confirmed that when an SNP influences one trait through another trait, the conditional analysis is more reliable than the marginal analysis to detect the mediating effects, thus distinguishing direct and indirect effects.

We applied our analysis to the BCX data for the subjects of European ancestry. Among the 25 SNPs that were both marginally significant for RBC and HCT, 10 of them became insignificant for RBC after adjusting for HCT. This may suggest these SNPs affect RBC through HCT, though it may be also due to reduced power in conditional analysis. We also applied our analysis to the GIANT data. We found that the joint models could reduce the number of significant SNPs, and more importantly, possibly better distinguish sex-specific associations.

We assume that all the traits were collected from the same set of subjects, which may not hold in practice due to missing values, as often shown in varying samples sizes for trait-specific summary statistics in GWAS data. If this assumption is violated, we may obtain estimates with bias, whose degree depends on the proportion of non-overlapping subjects. In the worst case of no-overlapping subjects among the multiple traits, a conditional analysis reduces to marginal analysis (because the traits would be estimated to be uncorrelated to each other).

In addition to conditional analysis of multiple traits to unravel specific pleiotropic effects, our proposed method can also be applied to conditional analysis of both multiple traits and multiple SNPs. We expect that our proposed method will be useful in future fine mapping studies.

The proposed method is implemented as an R function available at https://github.com/yangq001/conditional.

Acknowledgments

We thank the reviewers for constructive and helpful comments. This research was supported by NIH grants R01GM113250, R01HL105397 and R01HL116720, and by the Minnesota Supercomputing Institute at University of Minnesota.

References

Chami N, Chen MH, Slater AJ, Eicher JD, Lettre G. Exome genotyping identifies pleiotropic variants associated with red blood cell traits. Am J Hum Genet. 2016; 99:8–21. [PubMed: 27346685]

Cichonska A, Rousu J, Marttinen P, Pirinen M. metaCCA: Summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. Bioinformatics. 2016; 32(13):1981–1989. [PubMed: 27153689]

He Q, Avery CL, Lin DY. A general framework for association tests with multivariate traits in large-scale genomics studies. Genet Epidemiol. 2013; 37(8):759–767. [PubMed: 24227293]

Jiang Y, Li N, Zhang H. Identifying Genetic Variants for Addiction via Propensity Score Adjusted Generalized Kendall's Tau. J Am Stat Assoc. 2014; 109(507):905–930. [PubMed: 25382885]

Kim J, Bai Y, Pan W. An adaptive association test for multiple phenotypes with GWAS summary statistics. Genet Epidemiol. 2015; 39:651–663. [PubMed: 26493956]

Kim J, Zhang Y, Pan W. Alzheimer's Disease Neuroimaging Initiative. Powerful and Adaptive Testing for Multi-trait and Multi-SNP Associations with GWAS and Sequencing Data. Genetics. 2016; 203(2):715–731. [PubMed: 27075728]

Kwak I, Pan W. Adaptive gene- and pathway-trait association testing with gwas summary statistics. Bioinformatics. 2016; 32(8):1178–1184. [PubMed: 26656570]

Kwak I, Pan W. Gene- and pathway-based association tests for multiple traits with GWAS summary statistics. Bioinformatics. 2017; 33(1):64–71. [PubMed: 27592708]

Li R, Tsaih S-W, Shockley K, Stylianou IM, Wergedal J, Paigen B, Churchill GA. Structural Model Analysis of Multiple Quantitative Traits. PLoS Genet. 2006; 2(7):e114. [PubMed: 16848643]

Majumdar A, Haldar T, Witte JS. Determining Which Phenotypes Underlie a Pleiotropic Signal. Genet Epidemiol. 2016; 40(5):366–81. [PubMed: 27238845]

Randall JC, Winkler TW, Kutalik Z, Berndt SI, Jackson AU, Monda KL, Kilpeläinen TO, Esko T, Mägi R, Li S, Heid IM. Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. PLoS Genet. 2013; 9:e1003500. [PubMed: 23754948]

Schaid DJ, Tong X, Larrabee B, Kennedy RB, Poland GA, Sinnwell JP. Statistical Methods for Testing Genetic Pleiotropy. Genetics. 2016; 204(2):483–497. [PubMed: 27527515]

Stephens M. A unified framework for association analysis with multiple related phenotypes. PLoS One. 2013; 8(7):e65245. [PubMed: 23861737]

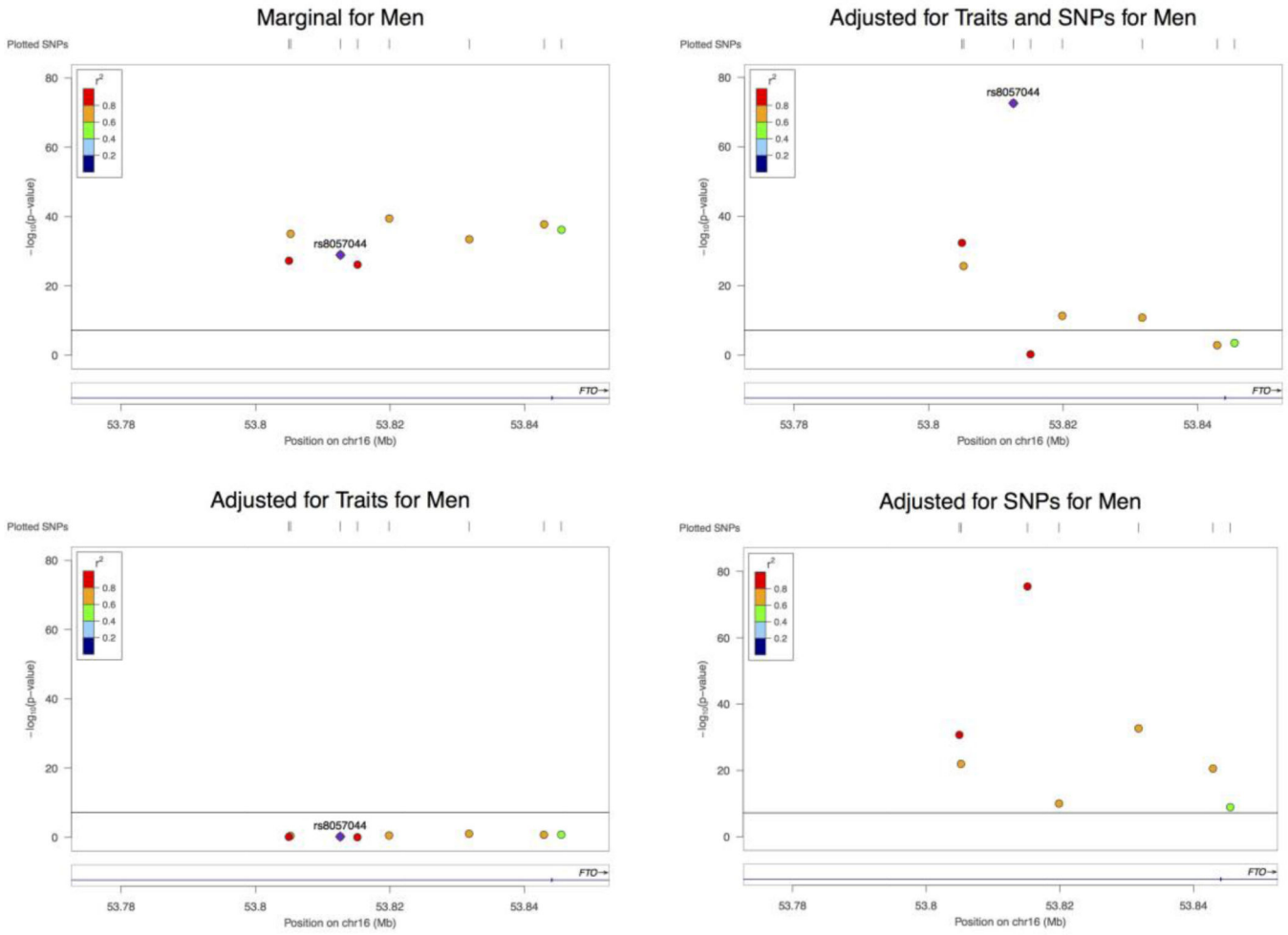
The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526:68–74. [PubMed: 26432245]

Wang Z, Sha Q, Zhang S. Joint Analysis of Multiple Traits Using "Optimal" Maximum Heritability Test. PLoS One. 2016; 11(3):e0150975. [PubMed: 26950849]

Yang J, Ferreira T, Morris AP, Medland SE, Visscher PM. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012; 44(4):369–375. [PubMed: 22426310]

Zhang Y, Xu Z, Shen X, Pan W. Alzheimer's Disease Neuroimaging Initiative. Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. Neuroimage. 2014; 96:309–325. [PubMed: 24704269]

Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N, Smith JA, Yanek LR, Sun YV, Edwards TL, Chen W, Nalls M, Fox E, Sale M, Bottinger E, Rotimi C, COGENT BP Consortium. Liu Y, McKnight B, Liu K, Arnett DK, Chakravati A, Cooper RS, Redline S. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. Am J Hum Genet. 2015; 96(1):21–36. [PubMed: 25500260]



The black horizontal lines at -$\log_{10}$p=7.3 indicate the genome-wide significance cut-off (p=5e-8).

Figure 1.

The effect of each SNP on BMI in different models (Men)

The black horizontal lines at −$\log_{10}$p=7.3 indicate the genome-wide significance cut-off (p=5e-8).

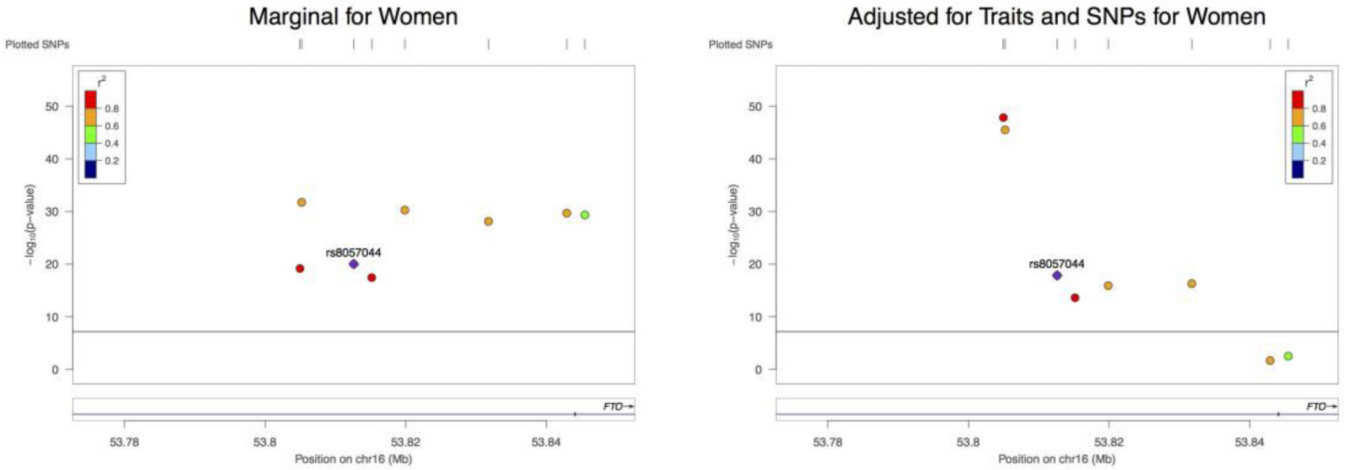


Figure 2.
The effect of each SNP on BMI in different models (Women)

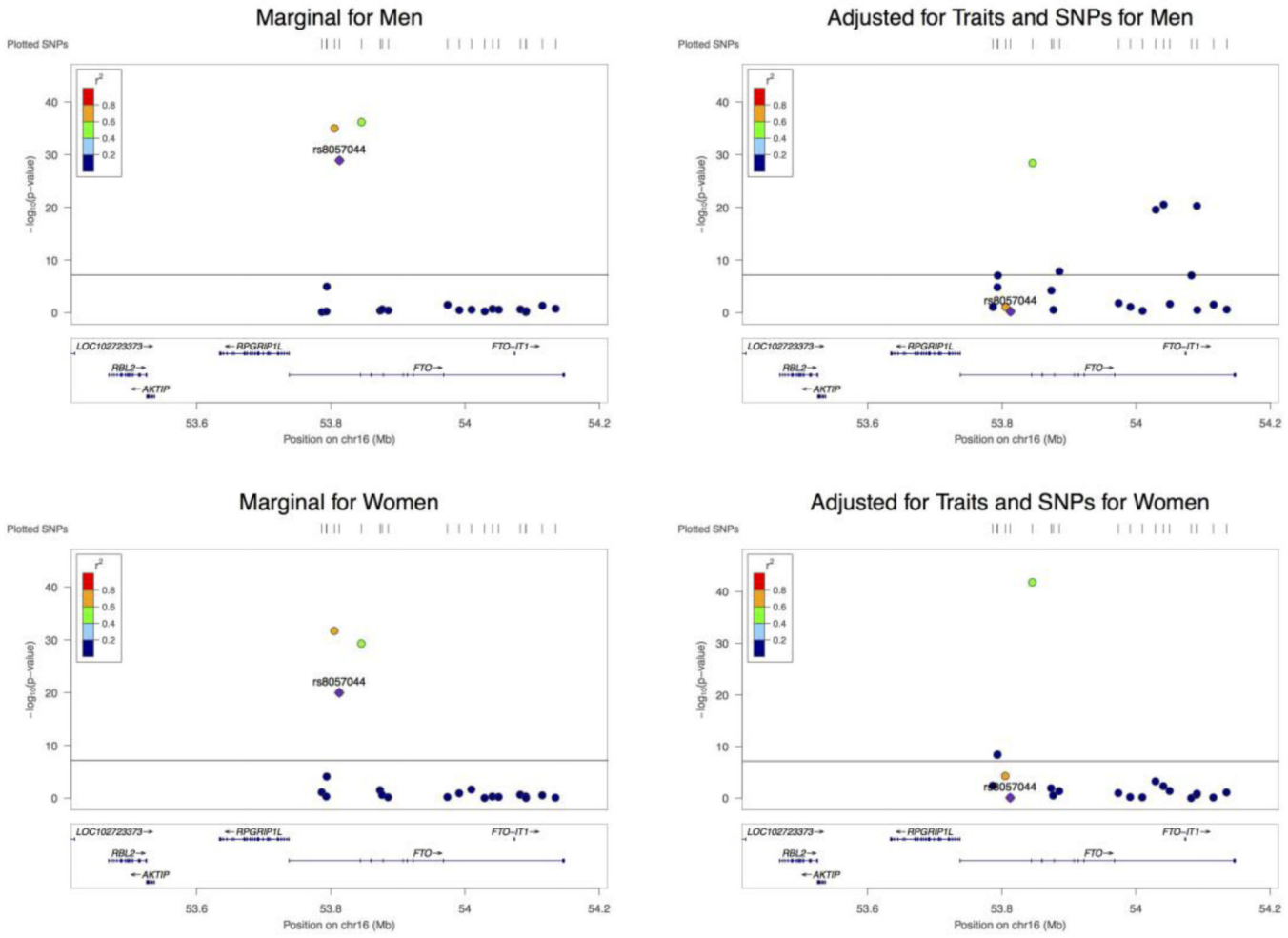


Figure 3.
The effect of each SNP on BMI in different models

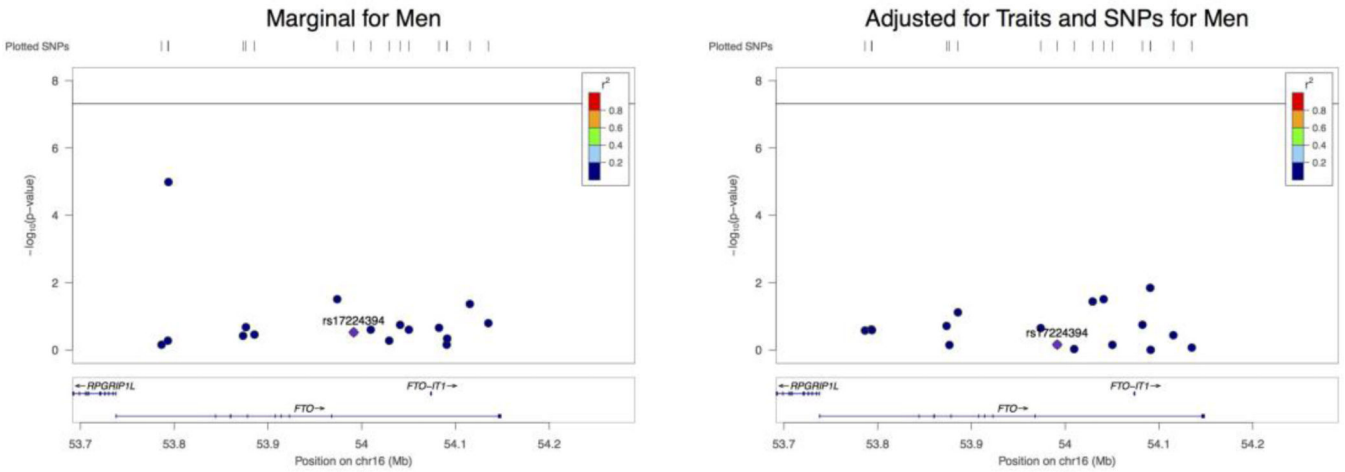


Figure 4.
The effect of each SNP on BMI in different models (Men)

Table 1

Estimated coefficients, standard errors and p-values when all subjects are overlapped

$f_0 = 0.1$, $(\beta_2, b_1, b_2) = (1, 0.8, 1.2)$, $n = 500$						
	$\widehat{b_1}$	s.e.	P	$\widehat{b_2}$	s.e.	P
summary	0.89	0.155	8e-9	1.20	0.044	2e-167
individual	0.90	0.157	2e-8	1.19	0.044	2e-98
$f_0 = 0.1$, $(\beta_2, b_1, b_2) = (1, 0.8, 1.2)$, $n = 2000$						
	$\widehat{b_1}$	s.e.	P	$\widehat{b_2}$	s.e.	P
summary	0.78	0.078	2e-23	1.16	0.022	0
individual	0.76	0.077	9e-23	1.17	0.022	0
$f_0 = 0.1$, $(\beta_2, b_1, b_2) = (0.5, 0, 0.5)$, $n = 2000$						
	$\widehat{b_1}$	s.e.	P	$\widehat{b_2}$	s.e.	P
summary	−0.06	0.074	0.45	0.49	0.022	3e-110
individual	−0.05	0.074	0.51	0.50	0.022	<2e-16

As shown in Table 1, the results of the two approaches were quite close, confirming the effectiveness of our proposed method.

Table 2

Estimated coefficients, standard errors and p-values when non-overlapping subjects exist for the two traits

$f_0 = 0.1$, $(\beta_2, b_1, b_2) = (1, 0.8, 1.2)$, $n = 500$, $n_1 = 400$, $n_2 = 500$				
	Mean($\widehat{b_1}$)	SD	Mean($\widehat{b_2}$)	SD
summary	0.82	0.22	1.18	0.09
individual	0.79	0.18	1.21	0.05
marginal	2.07	0.25	2.03	0.23
$f_0 = 0.1$, $(\beta_2, b_1, b_2) = (1, 0.8, 1.2)$, $n = 500$, $n_1 = 500$, $n_2 = 400$				
	Mean($\widehat{b_1}$)	SD	Mean($\widehat{b_2}$)	SD
summary	1.07	0.20	0.94	0.07
individual	0.80	0.18	1.21	0.05
marginal	2.03	0.23	2.01	0.25
$f_0 = 0.1$, $(\beta_2, b_1, b_2) = (1, 0.8, 1.2)$, $n = 500$, $n_1 = 400$, $n_2 = 400$				
	Mean($\widehat{b_1}$)	SD	Mean($\widehat{b_2}$)	SD
summary	1.16	0.26	0.84	0.10
individual	0.79	0.21	1.21	0.06
marginal	2.07	0.25	2.01	0.25

As shown in Table 2, the presence of non-overlapping subjects did affect the results of our method. The main reason is due to the poor approximation $\text{cor}(\mathbf{Y}_1, \mathbf{Y}_2) \approx \text{cor}(\mathbf{Z}_1, \mathbf{Z}_2)$ in this case. Table 3 shows the averages of the estimated correlations based on 20 samples for varying degrees of overlapping subjects. We generated one set of $(\mathbf{Y}_1, \mathbf{Y}_2)$, and then simulated $(\mathbf{Z}_1, \mathbf{Z}_2)$ 20 times for each set-up. It is clear that as the proportion of the non-overlapping subjects increased, the bias of the correlation estimates also increased.

Table 3

The correlation between Y_1, Y_2 and Z_1, Z_2 under each setting

$n = 500$	$n_1 = 500, n_2 = 500$	$n_1 = 400, n_2 = 500$	$n_1 = 500, n_2 = 400$	$n_1 = 400, n_2 = 400$
cor($\mathbf{Y}_1$, $\mathbf{Y}_2$)		0.77		
cor($\mathbf{Z}_1$, $\mathbf{Z}_2$)	0.77	0.71	0.69	0.59

Since the estimated correlation between $\mathbf{Y}1$ and $\mathbf{Y}2$ was under-biased, $\widehat{b_2}$ became underestimated. Meanwhile, $\widehat{b_1}$ was inflated. Nevertheless, the estimates from our conditional analysis were still better than those from marginal analysis (i.e. closer to the true parameters in the conditional model).

Table 4

Estimated coefficients, standard errors and p-values

SNP	marginal	s.e.	conditional	s.e.	"true"
rs10852521	−0.027	0.0055	−0.171	0.0047	−0.168
rs11075985	0.093	0.0051	0.134	0.0040	0.118
rs11075987	−0.019	0.0055	−0.163	0.0031	−0.192
rs11075989	0.057	0.0054	−0.043	0.0035	−0.050
rs11642841	0.070	0.0053	0.026	0.0029	0.006
rs12149832	0.084	0.0052	0.085	0.0032	0.068
rs8057044	0.106	0.0050	0.278	0.0031	0.304
rs9922619	0.042	0.0055	−0.092	0.0033	−0.076

The marginal models are BMI~SNP$_i$. The conditional model is BMI~HIP + WC + SNP$_1$ + SNP$_2$ + ⋯ + SNP$_8$. The true model was generated using the original GWAS data. The covariance for SNPs was estimated from the 381 European subjects from the 1000 Genomes Project data (The 1000 Genomes Project Consortium 2015). When building the conditional model for simulated data, we multiplied non-diagonal elements of the variance-covariance matrix by 0.8, since the same adjustment was made to generate these data. 2000 independent null SNPs with MAF $\sim$U [0.1, 0.5] were simulated to estimate the correlation between traits. λ = 0.01.

Table 5

Marginal and conditional effect on RBC

SNP	Chr	gene	MAF	Marginal			Conditional (on HCT)		
				$\widehat{\beta_1}$	s.e.	P	$\widehat{b_1}$	s.e.	P
rs218237	4		0.1143	−0.106	0.009	9e-33*	−0.075	0.008	1e-21*
rs2293767	7	ZAN	0.3012	0.066	0.007	6e-22*	0.044	0.006	3e-13*
rs10247980	7	ZAN	0.33	0.063	0.007	1e-20*	0.042	0.006	2e-12*
rs11611647	12		0.2227	−0.049	0.007	1e-11*	−0.030	0.006	2.0e-6
rs837763	16		0.5676	−0.035	0.006	3e-8*	−0.015	0.006	0.006
rs10495928	2	PRKCE	0.3171	−0.057	0.006	2e-19*	−0.017	0.006	0.003

Table 6

Estimated correlation between traits using Z_k's and the range of sample size for each trait (Upper: Men; Lower: Women)

Correlation	BMI (221, 58665)	HIP (118, 32851)	WC (118, 38313)	Weight (113, 58351)
BMI (124, 67959)	1	0.54	0.70	0.84
HIP (120, 40358)	0.63	1	0.68	0.61
WC (120, 47322)	0.72	0.70	1	0.70
Weight (125, 67594)	0.89	0.66	0.72	1

Table 7

Effective sample size for each trait-SNP pair (men)

SNP	BMI	HIP	WC	Weight
rs10852521	58612.9	32844.1	38299.9	58317.5
rs11075985	58612.1	32843.3	38299.4	58316.7
rs11075987	58615.9	32846.5	38302.5	58320.5
rs11075989	58618.1	32848.3	38304.2	58322.7
rs11642841	58604.5	32841.8	38297.5	58309.1
rs12149832	58597.6	32838.7	38294.8	58302.2
rs8057044	58556.0	32791.5	38247.5	58260.7
rs9922619	58615.8	32846.2	38302.3	58320.4

The (effective) sample size did not vary much among these SNPs for a given trait, but it did vary across the traits, suggesting at least about 1/4 of the subjects were not overlapped, thus cautions are needed with regard to possible biases of the parameter estimates.